D332–D333 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh138

dictyBase: a new *Dictyostelium discoideum* genome database

Lisa Kreppel, Petra Fey[1], Pascale Gaudet[1], Eric Just[1], Warren A. Kibbe[1], Rex L. Chisholm[1] and Alan R. Kimmel*

Laboratory of Cellular and Developmental Biology (50/3351), National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-8028, USA and [1]Center for Genetic Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

Received October 20, 2003; Revised and Accepted October 28, 2003

ABSTRACT

***Dictyostelium discoideum* is a powerful and genetically tractable model system used for the study of numerous cellular molecular mechanisms including chemotaxis, phagocytosis and signal transduction. The past 2 years have seen a significant expansion in the scope and accessibility of online resources for *Dictyostelium*. Recent advances have focused on the development of a new comprehensive online resource called dictyBase (http://dictybase.org). This database not only provides access to genomic data including functional annotation of genes, gene products and chromosomal mapping, but also to extensive biological information such as mutant phenotypes and corresponding reference material. In conjunction with additional sites (http://genome.imb-jena.de/dictyostelium/, http://dictyensembl.bioch.bcm.tmc.edu and http://www.sanger.ac.uk/Projects/D_discoideum/) from the genome sequencing and assembly centers, these improvements have expanded the scope of the *Dictyostelium* databases making them accessible and useful to any researcher interested in comparative and functional genomics in metazoan organisms.**

DICTYBASE: AN ONLINE INFORMATICS RESOURCE FOR *DICTYOSTELIUM DISCOIDEUM*

dictyBase [R. Chisholm, P. Fey, P. Gaudet, E. Just and W. Kibbe, Northwestern University Medical School (http://dictybase.org)] is a newly designed database (1) that integrates all currently available information on *Dictyostelium discoideum*, archiving genomic data, protocols, phenotypic information on mutant strains, images of numerous cellular processes, the Franke *Dictyostelium* Reference Library (http://dictybase.org/reference_database/index.html)], which compiles all *Dictyostelium* references in several downloadable formats, and DictyNews, a searchable newsletter. There are also site links to laboratories that use *Dictyostelium* in their research, the *Dictyostelium* Genome Centers (http://genome. imb-jena.de/dictyostelium/, http://dictyensembl.bioch.bcm.tmc.edu and http://www.sanger.ac.uk/Projects/D_discoideum/) and the *Dictyostelium* cDNA Project of Japan (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html). The schema and Perl code for dictyBase are a modified version of the *Saccharomyces* Genome Database (2,3).

dictyBase facilitates full compilation of *D.discoideum* genomic data. The collective chromosomal sequence of *Dictyostelium* has been estimated at 34 Mb distributed among six chromosomes, of ~4–6 Mb each. Chromosome 2, which comprises ~25% of the genome, is sequenced and assembled (1) to near completion (4) in 50 contigs. Chromosome 1 (4.7 Mb) is assembled into four contigs and chromosome 6 is represented by 15 contigs. Sequencing is largely complete, but assembly continues on the remainder of the genome. The Sanger Institute has compiled an ~3500 contig set that represents a whole-genome assembly draft (http://www.sanger.ac.uk/Projects/D_discoideum/genomic_sequence.shtml, and sites therein). Although much of the genomic data are either unpublished or unavailable through GenBank or related databases, all of the *Dictyostelium* genome centers deposit 'finished' sequences in dictyBase. The consortium generously provides full access for all their analyses (for guidelines, see http://www.sanger.ac.uk/Projects/D_discoideum/data_release.shtml). These data are complemented by sequences of ~75 000 developmentally staged, full- and partial-length cDNAs at the *Dictyostelium* cDNA Project.

To promote access and analyses of the extensive sequence data available from all the centers, dictyBase completely integrates genomic and EST assembly and curation. Each gene has its own locus page, which links all significant information including sequence data, chromosomal map, alternative names and aliases, protein information, gene ontology (GO) annotations, relevant references and a BLAST (5) server.

The chromosomal maps utilize the 'Generic Genome Browser' (Gbrowse; http://www.gmod.org/ggb/index.shtml), implemented from the Generic Model Organism Database Construction Set. Gbrowse visually displays features in the database based on their relative location on a chromosome. Users can view graphical displays of any chromosomal region, by zooming and centering, by name searching or by position. Clicking a feature in Gbrowse brings up the locus page with all known annotations. In addition, users can customize their

*To whom correspondence should be addressed. Tel: +1 301 496 3016; Fax: +1 310 496 5239; Email: ark1@helix.nih.gov

display, view restriction sites and produce any sequence in a variety of formats. They can also find alignments of genes or gene models with ESTs and contigs. Currently, Gbrowse is most useful for the assembled chromosomes 1, 2 and 6.

dictyBase is updated and improved by continuous manual curation of all genomic data. An important aspect of curation is the use of GO (http://www.geneontology.org/), a controlled vocabulary for the description of molecular function, biological process and cellular component of gene products. Programs that scan all assembled contigs predict ~13 000 ORFs [http://dicty.sdsc.edu/annot-020303.html and http://dicty.sdsc.edu/ (version 3; N. Iranfar, W.F. Loomis and T.B.K. Reddy, University of California, San Diego)]. Of these, ~1800 can be linked (P. Bourne, W. Li and V. Reyes, San Diego Supercomputer Center) to the Protein Data Bank of 3D structures (6,7). Gene prediction programs are still subject to enormous inaccuracies so manual verification of each gene is essential. However, ~6400 ORFs are represented as unique ESTs or full-length cDNAs (8) and, thus, are largely confirmed.

The GO terms are used by most model organism databases and provide a uniform platform to facilitate comparison between genes of different model organisms based on a common ontological language. Whenever possible gene names in dictyBase will follow the proposed nomenclature for *Dictyostelium* (http://dictybase.org/Nomenclatureproposal.htm), and researchers can reserve names for genes that are works in progress. Both Demerec (9) and non-Demerec names are linked to genes allowing them to be located by searches on these names. External links include all Entrez Nucleotide and Protein records.

Finally, dictyBase will provide direct and searchable access to the resources of the NIH-supported *Dictyostelium* Stock Center, which is maintained by J. Franke and R. Kessin (Columbia University). The facility is a central repository for *Dictyostelium* strains, mutants, and plasmid constructs and is intended to preserve and to freely distribute these essential and vulnerable materials. Requests will be produced directly from the pages describing each stain, mutant or plasmid.

ACKNOWLEDGEMENTS

We are indebted to all of our colleagues who focus much of their research efforts on the analyses of *Dictyostelium*, with particular mention of the International Genomic Sequencing Consortium [The University of Cologne (A. Noegel and L. Eichinger), GSC Jena (G. Gloeckner, M. Platzer and A. Rosenethal), The Baylor College of Medicine (A. Kuspa, R. Sucgang and R. Gibbs), The Sanger Institute (B. Barrell, M.-A. Rajandream and M. Quail)] and the Japanese *Dictyostelium* cDNA Group [the universities of Tsukuba (Y. Tanaka, H. Urushihara, T. Morio, M. Katoh and H. Kuwayama), Hokkaido (H. Ochiai and T. Saito) and Osaka (M. Maeda)]. We wish to specifically thank Drs J. Brzostowski, F. Comer, P. Dyck, T. Khurana, S. Merchant, C. Parent and D. Rosel for their many helpful discussions and suggestions. dictyBase is supported by a grant from the National Institute of General Medical Sciences (GM064426).

REFERENCES

1. Kreppel,L. and Kimmel,A.R. (2002) Genomic database resources for *Dictyostelium discoideum*. *Nucleic Acids Res.*, **30**, 84–86.
2. Issel-Tarver,L., Christie,K.R., Dolinski,K., Andrada,R., Balakrishnan,R., Ball,C.A., Binkley,G., Dong,S., Dwight,S.S., Fisk,D.G. *et al.* (2002) *Saccharomyces* Genome Database. *Methods Enzymol.*, **350**, 329–346.
3. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroede,M., Sherlock,G. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
4. Glockner,G., Eichinger,L., Szafranski,K., Pachebat,J.A., Bankier,A.T., Dear,P.H., Lehmann,R., Baumgart,C., Parra,G., Abril,J.F. *et al.* and the *Dictyostelium* Genome Sequencing Consortium (2002) Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature*, **418**, 79–85.
5. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
8. Urushihara,H. (2002) Functional genomics of the social amoebae, *Dictyostelium discoideum*. *Mol. Cell*, **13**, 1–4.
9. Demerec,M., Adelberg,E.A., Clark,A.J. and Hartma,P.E., (1966) A proposal for a uniform nomenclature in bacterial genetics. *Genetics*, **54**, 61–76.